Animating the Past: Reconstruct Trilobite via Video Generation

Xiaoran Wu*
AI Lab
Yishi Inc.
Hangzhou, China
wuxr18@tsinghua.org.cn

Zien Huang*
The International Department
Experimental High School
Attached to Beijing Normal University
keane.huangzien@gamil.com

Chonghan Yu
School of Ocean Sciences
China University of Geoscience
Beijing, China
yu_chonghan@foxmail.com

Fig. 1: We design and train the first text-to-video framework that automatically learn to refine the prompts to generate visually realistic trilobites adhering closely to pronounced and authentic trilobite characteristics in more fluid and lifelike videos. The first prompting image is courtesy of [1].

***Abstract*—Paleontology, the study of past life, fundamentally relies on fossils to reconstruct ancient ecosystems and understand evolutionary dynamics. Trilobites, as an important group of extinct marine arthropods, offer valuable insights into Paleozoic environments through their well-preserved fossil records. Reconstructing trilobite behaviour from static fossils will set new standards for dynamic reconstructions in scientific research and education. Despite the potential, current computational methods for this purpose like text-to-video (T2V) face significant challenges, such as maintaining visual realism and consistency, which hinder their application in science contexts. To overcome these obstacles, we introduce an automatic T2V prompt learning method. Within this framework, prompts for a fine-tuned video generation model are generated by a large language model, which is trained using rewards that quantify the visual realism and smoothness of the generated video. The fine-tuning of the video generation model, along with the reward calculations make use of a collected dataset of 9,088 *Eoredlichia intermedia* fossil images, which provides a common representative of visual details of all class of trilobites. Qualitative and quantitative experiments show that our method can generate trilobite videos with significantly higher visual realism compared to powerful baselines, promising to boost both scientific understanding and public engagement.**

***Index Terms*—Trilobite, *Eoredlichia intermedia*, Text-to-Video, Multimodal Large Language Model, Learning from Human Feedback**

I. Introduction

Paleontology, the study of prehistoric life, relies heavily on the fossil record to reconstruct past ecosystems, understand evolutionary processes, and decipher extinct organisms' biology [2], [3]. As an extinct group of marine arthropods, trilobites are among the most iconic and well-studied fossils [2], [4], [5], providing critical insights into Paleozoic ecosystems. Reconstructing the behavior and locomotion of trilobites is of great research and educational interests [1]. Such dynamic reconstructions help in formulating hypotheses about trilobites' living environments and the functional morphology and ecological roles of these ancient creatures [4]–[7]. Furthermore, from the educational aspect, reconstruction provides a tangible visualization of trilobite appearance and behavior, thus bridging the gap between abstract scientific knowledge and public understanding [8].

Despite the abundance in the trilobite fossil record, reconstructing their behavior and movement remains a challenge, primarily due to the limitations inherent in fossil remains, such as their static nature. Fortunately, recent advancements in generative artificial intelligence (AI) and computational techniques provide new opportunities to address these challenges [9]–[13]. Integrating AI into paleontological research not only showcases the potential of extending machine learning into a field of natural research that AI has not studied extensively before [14] but also hopefully can enhance our understanding of the trilobite ethology and shed new light on its study.

Among generative AI techniques, video generation [1], [15], [16] techniques in particularly suitable for simulating trilobite movement in a dynamic, visually engaging manner. However, the current methods of video generation encounter several challenges that hinder their application to paleontological

*These authors contributed equally to this work.

reconstructions. Primarily, as demonstrated in our qualitative studies, existing methods struggle with maintaining the realism of the depicted trilobites, with the creatures appearing unrealistic or oddly shaped [3]. This lack of realism significantly detract from the viewer's engagement and reduce the educational and research value of the visualizations. Moreover, the consistency of generated videos often falls short, with noticeable discrepancies between consecutive frames [17], [18]. Such inconsistencies are particularly problematic in longer sequences, leading to choppy transitions that disrupt the fluid simulation of trilobite movement.

To tackle these issues, we propose a novel approach that embeds the evaluation of trilobite realism and video smoothness directly into the video generation workflow. Our solution leverages diffusion models [19]–[22], which have demonstrated impressive capabilities in producing realistic images and videos from textual descriptions. We employ these models to create a series of animated segments that capture various aspects of trilobite movement, guided by descriptive prompts generated by a large language model (LLM) [23]–[25]. The cornerstone of our method involves assessing the smoothness of transitions and the accuracy of the trilobite appearance in these animations, compared against a curated collection of trilobite fossil images. This assessment acts as a feedback mechanism to fine-tune the LLM that generates prompts for the text-to-animation model [26]–[28], enhancing the fidelity of animations. The objective is dual: to produce animations that not only accurately depict trilobite appearance and movement but also ensure seamless transitions, adding a layer of complexity to the model's training but crucial for high-quality video output.

In summary, our methodology encompasses several stages: initially, we generate basic animated segments from a set list of LLM-generated prompts. These segments are then pieced together, and the composite video is evaluated for the quality of its transitions and the realism of its content. The evaluation results are used as reward signals to update the LLM with preference optimization [29], [30] to refine the animations. This cycle of generation, evaluation, and enhancement is repeated until the video meets our criteria for smoothness and realism.

We comprehensively evaluate our method both qualitatively and quantitatively against state-of-art text-to-animate and text-to-video academic research [27] and commercial tools [31], [32]. The results show clear advances in paleontological visualization in terms of content realism and video continuity. Furthermore, we provide ablation studies to show the contribution of each component in our learning framework. We hope that this pioneering integration of technology and paleontology makes significant contributions to the field of synthetic media generation and opens new pathways for visualizing and understanding prehistoric entities and exploring ancient life.

II. Related Work

Our method of training the Large Language Model (LLM) that generates prompts is related to Reinforcement Learning from Human Feedback (RLHF), an important technique to ensure that LLM outputs align with human preferences [33]–[35]. Here in our work, the counterpart of human preference is defined by metrics regarding content realism and video continuity. Typically, RLHF initially learns a reward model (RM) [36] from human preferences and then optimizes the supervised fine-tuned LLM model with reinforcement learning algorithms, e.g., PPO [37], to maximize the cumulative rewards from the RM. However, training the reward model is time-consuming and computation-intensive [36]. Direct Preference Optimaztion (DPO) [29] avoids training the reward model by directly aligning LLMs to best satisfy human preferences using a simple classification objective. The recently proposed KTO [30] extends DPO by maximizing utility functions derived from prospect theory [38] for accurate human utility modelling. In this work, to achieve better stability and robustness, we utilize the calculated realism and continuity rewards to order different LLM outputs (prompts to the animation generation model) and use KTO for preference optimization.

Video generation methods like Tune-a-Video [13] extend text-to-image (T2I) models to generate multiple images simultaneously by incorporating a tailored spatio-temporal attention mechanism and an efficient one-shot tuning strategy to learn continuous motion among generated images. Text2Video-Zero [12] proposes a cost-effective approach that requires no training or optimization by leveraging the capabilities of existing T2I synthesis methods, adapting them for the video generation domain. CogVideo [16] proposes a multi-frame-rate hierarchical training strategy to better align text and video clips on large-scale text-video datasets. Furthermore, commercial video generation tools are setting significant benchmarks. In this paper, we empirically compare our method against Pika [31] and Gen3 [32] for evaluation. Text-to-Animation (T2A) is another approach in video generation that extends pre-trained T2I models by incorporating temporal structures [26], [27]. AnimateDiff [27] introduces a plug-and-play motion module that enables the training of T2A models without the need for model-specific tuning. In this paper, we employ the T2A method to generate trilobite animations from user prompts, but the focus is on how to enhance the temporal coherence and content realism.

We now introduce the preliminaries of the RLHF and T2A techniques based on which we develop our method.

III. Preliminaries

RLHF. The training of modern Large Language Models (LLMs) involves three phases as outlined in [23], [33], [39], [40]. *(1) Pretraining:* This phase involves training an initial model π_0 on a large text corpus to optimize the prediction of the next token based on the preceding text [41]–[43]. *(2) Supervised Fine-tuning (SFT):* The model is further trained on task-specific data that generally includes targeted instructions and expected responses, to refine its utility for practical applications [44]. This fine-tuned model is noted as π_{ref}. *(3) RLHF:* This step uses a preference dataset $\mathcal{D}$ containing tuples

(x, y_w, y_l) where x is the input and y_w, y_l are the preferred and less preferred outputs, respectively [29], [34]. A Bradley-Terry model [45] is used here to calculate preferences:

$$p^*(y_w > y_l|x) = \sigma(r^*(x, y_w) - r^*(x, y_l)),$$

where σ is the logistic function. A reward model r_ϕ is trained by minimizing the negative log-likelihood of the preference data in the set $\mathcal{D}$ [33]:

$$\mathcal{L}_R(r_\phi) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[-\log\sigma(r_\phi(x, y_w) - r_\phi(x, y_l))].$$

To balance reward maximization with linguistic correctness, a KL divergence penalty is applied, preventing the model from deviating excessively from the reference model π_{ref}:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[r_\phi(x, y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(\cdot|x)||\pi_{\text{ref}}(\cdot|x)\right].$$

The objective, non-differentiable in nature, requires an RL approach like PPO [37] for optimization.

The computational demands and instability of training the reward model r_ϕ have led to the development of Direct Preference Optimization (DPO) [29], providing a stable alternative that trains directly on preference pairs and with similar optimal policy convergence performance:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]. \quad (1)$$

T2A. One approach for short video generation is to animate a text-to-image (T2I) [46] model by integrating temporal dynamics. Following the methodologies outlined in related literature [16], [27], a batch of video data is represented as 5-dimensional tensors $x \in \mathbb{R}^{b\times c\times f\times h\times w}$. Here, b denotes the batch axis, f represents the frame-time axis, and c, h, and w are the channels, height, and width of each video frame, respectively. The text-to-animation process begins by encoding each frame of a video data batch $x_{1:f} \in \mathbb{R}^{b\times c\times f\times h\times w}$ into latent representations $z_{1:f,0}$ using a pre-trained auto-encoder. These representations are subsequently perturbed by noise as per the forward diffusion schedule [20], [47]:

$$z_{1:f,t} = \sqrt{\bar{\alpha}_t}z_{1:f,0} + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad t = 1, \ldots, T. \quad (2)$$

In this inflated model, the noisy latent representations, along with corresponding text prompts, serve as inputs for predicting the noise added during the diffusion process. The training objective for T2As can be formulated as:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x_{1:f}),y,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(z_{1:f,t}, t, \tau_\theta(y))\|_2^2\right]. \quad (3)$$

By inflating the model with the additional temporal axis [48], this formulation emphasizes the critical role of temporal coherence and dynamic content adaptation in generating animations from textual descriptions.

However, this approach may yield animations that are not only brief in duration but also suffer from less smooth transitions between frames and inconsistencies in object appearances across different frames. These issues primarily arise from the model's limitations in maintaining consistent motion patterns and visual quality throughout the sequence [17]. This challenge is particularly pronounced when the model attempts to interpolate complex dynamics, a task that demands high fidelity in temporal and spatial representations. The difficulty lies in the model's capacity to accurately generate and link successive frames where each must evolve naturally from its predecessor while adhering to the dynamics specified by the textual description.

IV. Method

In this section, we describe our method that addresses the challenge of maintaining motion smoothness and visual realism throughout the video sequences.

The proposed framework synergizes the power of a large language model (Script Writer) that generates prompts and a fine-tuned text-to-animation model (Video Generator). Our main technical novelty lies in the design of the optimization algorithm of Script Writer. In effect, we design a contextual bandit learning task for the Script Writer. The concept of contextual bandit is popular in the cutting-edge LLM research, such as direct preference optimization (DPO [29]).

For fine-tuning the Video Generator and training the Script Writer, we collect a set $\mathcal{R}$ of 9088 *Eoredlichia intermedia* fossil images, which include a large number of specimens covering different stages of individual development of *Eoredlichia intermedia* trilobites. The images are used to provide a common representative of the visual traits of all class of trilobites in order to enhance the visual details of the generated content. These real trilobite fossil images do not mean that the videos produced in this study can fully reproduce the real trilobite structure, but the use of these real fossil details can greatly supplement the scarcity and errors of trilobite images in the web footage, enhancing the trilobite structure and details in the videos.

A. Prompt and Video Initialization

We now describe the details of our method. As the first step, the Script Writer $\pi(\theta^0)$, where θ^0 is the initial parameters, is asked to generate an initial prompt y^0 for the text-to-animation model with the format

$$y^0 = (t_1 : y_1^0; t_2 : y_2^0; , \cdots, t_N, y_N^0), \quad (4)$$

where $y_n^0, n \in [N]$ is a textual description of the appearance and expected movement of a trilobite in the animation clip n, and $t_n, n \in [N]$ is the frame index from which the animation clip n will start in the final video.

This initial prompt y directs the text-to-animation diffusion model Video Generator to generate N initial animation clips $(c_1^0(y^0), \cdots, c_N^0(y^0))$ that are concatenated sequentially to get an initial video $z_{1:f}^0(y^0)$. Before generating this initial video, the Video Generator has been fine-tuned on the collected fossil image dateset $\mathcal{R}$ to enhance the model's ability to generate the detailed textures and structures observed in fossil images.

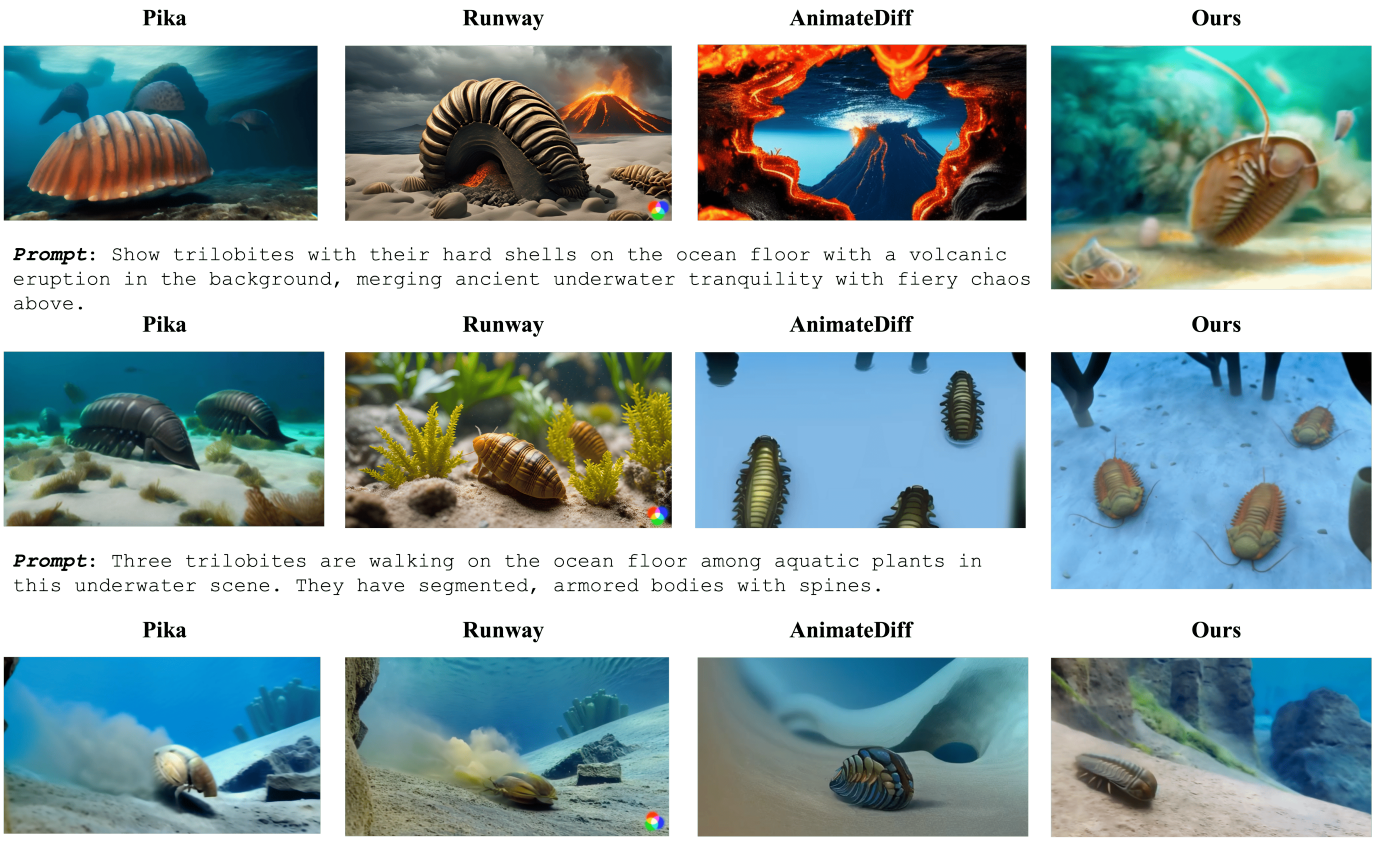

Fig. 2: Qualitative comparison of the generated videos from four different models: Pika [31], Runway [32], AnimateDiff [27], and ours. Our model significantly outperforms the others in generating trilobites with highly detailed morphological accuracy, realistic texturing, and appropriate environmental interactions. The prompting images in the first and second rows are courtesy of [1] and [49], respectively.

B. Reward Design for the SCRIPT WRITER

The second step is to design reward signals to train the SCRIPT WRITER with the hope that it can refine the initial prompt y^0 so that the resulting video has a better quality in terms of transition smoothness and visual realism.

To be specific, the reward for prompt y^0 is designed to be a summation of two components: $r(y^0) = r_s(y^0) + r_a(y^0)$, where r_s measures the smoothness of frame transition and r_a measures the visual realism of the trilobites in the generated video.

Smoothness of Frame Transition. To assess the smoothness of a video, we compute the Fréchet Inception Distance (FID) [50] between adjacent frames. Consider two frames $x_t \in \mathbb{R}^{c\times h\times w}$ and $x_{t+1} \in \mathbb{R}^{c\times h\times w}$, where c, h, and w represent the channel, height, and width of the frame, respectively. We first use a pre-trained InceptionV3 network [51] to extract the image features (pool3 layer) z_t and z_{t+1} given frames x_t and x_{t+1}, and then compute the FID score for consecutive frames by:

$$\text{FID}_t = ||z_t - z_{t+1}||^2. \tag{5}$$

After obtaining the FID scores for all consecutive frames, we get the transition smoothness reward $r_s(y^0) = -\sum_{t=1}^{f} \text{FID}_t$. For fine-grained control, the reward can be calculated for each clip separately:

$$r_s(y_n^0) = -\sum_{t=t_n}^{t_{n+1}} \text{FID}_t. \tag{6}$$

Visual Realism. To ensure scientific rigorousness, we compare the visual details of the generated content against real samples of trilobite fossils from $\mathcal{R}$. For a video consisting of a set of frames, we expect that no frame has trilobites with morphological details deviating significantly from realistic data. To this end, we design a max-min objective:

$$r_a(y_n^0) = -\max_{x\in[f]} \min_{r\in\mathcal{R}} D(x, r) \tag{7}$$

Here, D is a distance function that measures the morphological similarity between a generated trilobite and a reference image. We now explain the intuition of this similarity reward. The reference image set contains images of trilobite fossils from different growing stages, various sizes, different levels of

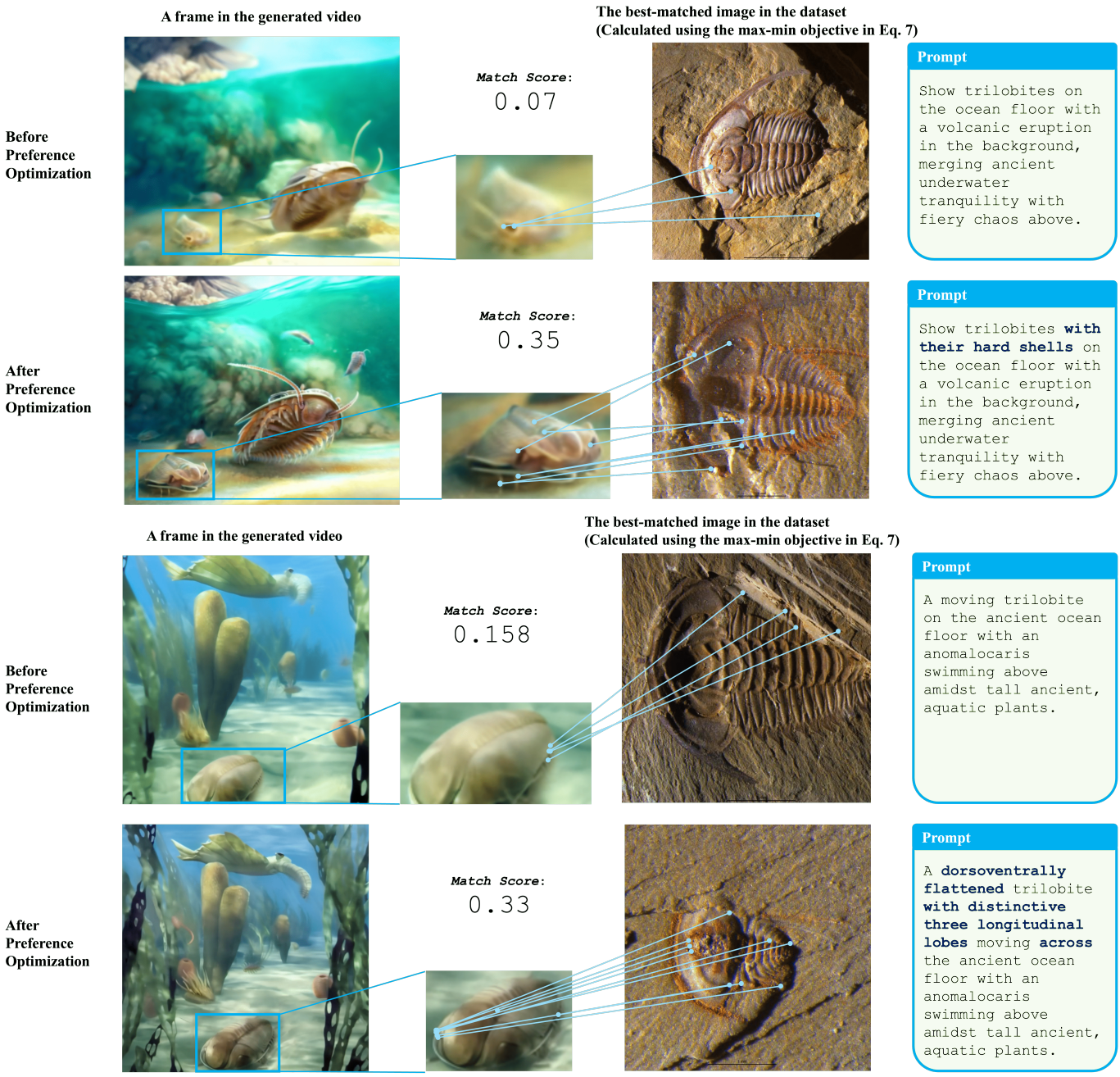

Fig. 3: The preference optimization of SCRIPT WRITER contributes to visual realism of the generated trilobites. What is **highlighted** represents the changes in prompts after preference optimization. We can see that **SCRIPT WRITER learns to improve the quality of generated videos by adding more descriptions about the trilobite morphological details**. The shown match score is the inverse of the distance from the most similar reference image ($1/\min_{r\in\mathcal{R}} D(x,r)$), which significantly improves after optimization. The first prompting image is courtesy of [1]. ***Please note**: The reference images are used to enhance the visual details of the generated content. We show the reference image with the highest match score calculated by the ORB detector. The connecting lines represent matching points with similar local image features, and do **not** mean that the trilobite in the video matches the trilobite in the real fossil image.*

preservation, and different geological periods. Therefore, we would have clear evidence that a generated trilobite is visually realistic if it is morphologically similar to at least one reference image. We capture this by the minimum operation in Eq. 7. Then we try to find the frame that is the most different from the reference set, minimizing which could guarantee that no frame deviates too much from the reference set.

In practice, we use the ORB (Oriented FAST and Rotated BRIEF (Binary Robust Independent Elementary Feature) [52]) detector as the distance function D. The ORB detector is a fast and efficient feature detection algorithm. It combines the FAST keypoint detector and the BRIEF descriptor, providing robust performance suitable for extracting morphological details. We then use BF (Brute-Force) [53], [54] method for matching the features. computes distances between every pair of descriptors, typically employing the Hamming distance for binary descriptors, as utilized in our case.

C. Training the SCRIPT WRITER

Having defined the reward signals, we now introduce the third step, the training of SCRIPT WRITER. We note that the rewards defined in the previous section are all negative, could be large in magnitude, and prone to noise, which indicates that these rewards may not be effective when used to train the SCRIPT WRITER LLM with algorithms like PPO [37], as they are sensitive to the specific values of the rewards. To tackle this problem, we propose ordering the prompts based on the rewards then applying preference optimization. Preference optimization has been proven its effectiveness in the broad literature of reinforcement learning from human feedback (RLHF) [33] and is more robust when reward values are noisy.

To be specific, we collect a training dataset $\mathcal{D}$ where each training sample contains a query x, a desirable generation y_d, and an undesirable generation y_u. Here, $r(y_d) > r(y_u)$. We use Kahneman-Tversky optimization (KTO) [30] to train the SCRIPT WRITER. Using λ_y to denote λ_D (λ_U) when y is desirable (undesirable), where λ_D and λ_U are two constants, the KTO loss is:

$$\mathcal{L}_{\text{KTO}}(\theta^0) = \mathbb{E}_{x,y\sim\mathcal{D}}[\lambda_y - v(x,y)] \tag{8}$$

where

$$r_{\theta^0}(x,y) = \log \frac{\pi_{\theta^0}(y|x)}{\pi_{\text{ref}}(y|x)}; \tag{9}$$

$$z_0 = \mathbb{E}_{x'\sim\mathcal{D}}[KL(\pi_{\theta^0}(y'|x') \parallel \pi_{\text{ref}}(y'|x'))]; \tag{10}$$

$$v(x,y) = \begin{cases} \lambda_D \sigma(\beta(r_{\theta^0}(x,y) - z_0)) & \text{if } y \sim y_d|x; \\ \lambda_U \sigma(\beta(z_0 - r_{\theta^0}(x,y))) & \text{if } y \sim y_u|x. \end{cases} \tag{11}$$

In the following section, we present examples showing how the KTO training improves the prompts.

After KTO training, SCRIPT WRITER parameters are updated to θ^1, which leads to updated prompts y^1. θ^1 can be further improved by running KTO given the new video generated by y^1. This process can be repeated until the new video is satisfactory enough.

V. EXPERIMENTS

In this section, we present experiments to test the effectiveness of our method. The experiments are designed to give both qualitative and quantitative evaluation of the generated videos. We compare against strong baselines, including previous work on text-to-animation (AnimateDiff [27]) and powerful commercial text-to-video tools, Pika Labs [31] and Gen-3 [32]. We also carry out ablation studies to show the separate contribution of SCRIPT WRITER training and VIDEO GENERATOR fine-tuning.

A. Qualitative Results I: Visual Realism

Compare with baselines. In Fig. 2, we showcase a qualitative comparison of trilobite renderings produced by different models. The objective of the comparison is to evaluate each model's ability to generate realistic trilobites in various dynamic backgrounds.

The first row shows a scene featuring trilobites on the ocean floor with a volcanic eruption in the background. The Pika model generates a trilobite with unrealistic segmentation. The Runway model shows a more realistic structure but lacks in capturing the authentic texture of trilobite exoskeletons. The AnimateDiff model produces an oversimplified trilobite, and the main part of the image features a volcano. In contrast, the trilobites generated by our model display intricate segmentation, realistic texturing, and coloration that blends well with the naturalistic ocean floor setting, making them the most lifelike.

The depiction in the second row includes three trilobites among aquatic plants on the seabed. The Pika model's trilobites are not similar to any known types of trilobites. Runway's versions show better integration with the background but are still somewhat artificial in appearance. The trilobites by AnimateDiff lack depth and detail in texturing. Our model, however, shows trilobites with precise, well-defined segmentation and natural colors that harmonize with the underwater environment, enhancing the realism of the scene.

The scene of the third row captures a single trilobite moving along the ocean floor with a focus on the interaction with the environment, such as sediment displacement. Pika's rendition again lacks realism in appearance. AnimateDiff's trilobite appears round. Meanwhile, Our model produces a realistic trilobite interacting with its surroundings, showing sediment displacement that suggests a natural weight and presence in the water.

In summary, our model outperforms in creating trilobites with realistic anatomical features, textural fidelity, and appropriate environmental interactions compared to the other models. This qualitative analysis underscores the ability of our method in generating video content that closely mirrors the true appearance of trilobites.

Compare with ablations. Two components in our method contribute to the visual realism of the generated trilobites: we first fine-tune the T2A model and then carry out preference optimization for SCRIPT WRITER. In Fig. 3, we show the influence of these two components. Specifically, we present

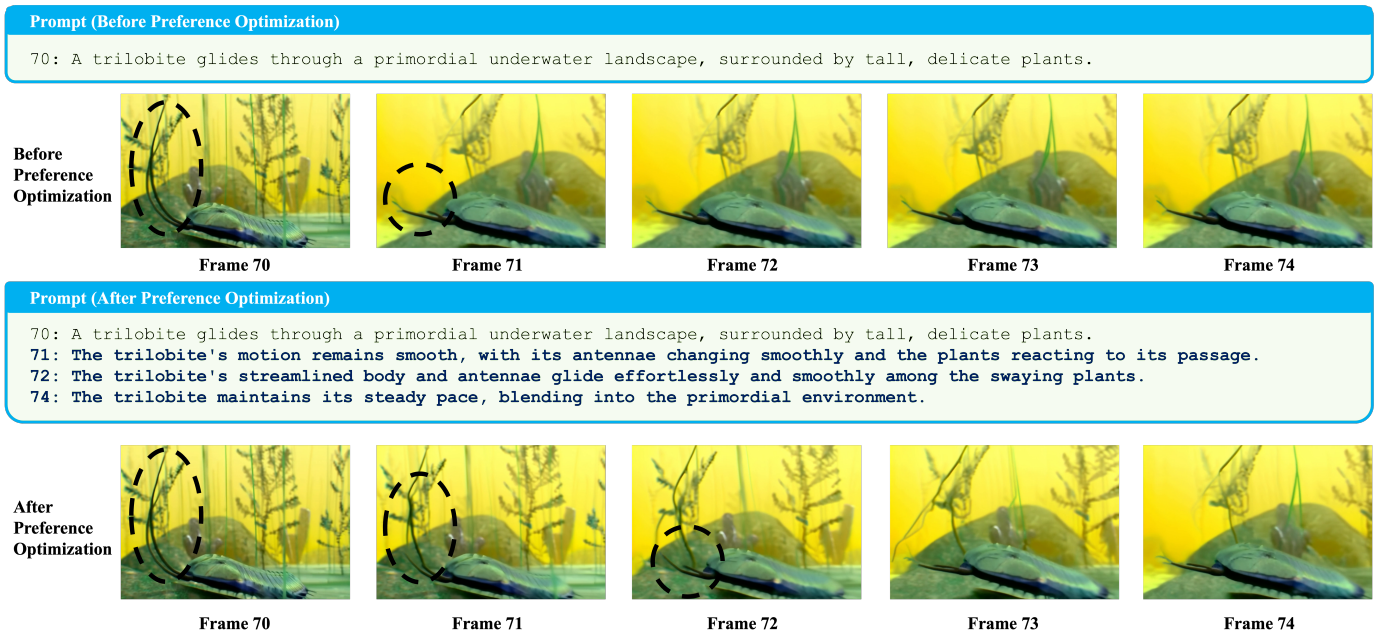

Fig. 4: A qualitative comparison before and after SCRIPT WRITER preference optimization, with a focus on the smoothness and continuity of the video. The SCRIPT WRITER learns to **add more prompts** to impact the smoothness of the resulting video. The prompting image is courtesy of [55].

two examples that demonstrate the results before and after preference optimization, with a focus on the visual quality of trilobite renderings in generated videos. Each example shows a frame from the video, accompanied by the most closely matching image from the dataset and the corresponding prompts.

In the first example (Fig. 3), before optimization, the video frame shows a trilobite on the ocean floor with a volcanic eruption in the background. The trilobite appears somewhat blended into the background, lacking distinct features, resulting in a low match score of 0.07. The prompt focuses on the general presence of trilobites amid a dynamic background. By contrast, after optimization, the optimized frame exhibits a trilobite with more emphasized and defined hard shells, enhancing its visibility and structural integrity against the complex background. This optimization is attributable to the updated prompt, which now specifically highlights the trilobite's hard shell. The match score significantly improves to 0.35, indicating a closer resemblance to the most similar reference image, which shows clearer and more detailed trilobite features.

In the second example (Fig. 3), before optimization, the original video frame captures a trilobite moving across the ocean floor with anomalocaris in the background. Initially, the trilobite lacks prominent distinguishing features, leading to a match score of 0.158. After preference optimization, the frame now shows the trilobite with enhanced distinguishing features, such as the longitudinal lobes and textural details, making it more realistic and akin to the reference image. Again, it is the updated prompt that leads to these changes, specifically by pointing out these features, contributing to a raised match score of 0.33.

In both cases, preference optimization led to adjustments in the model's rendering, focusing on enhancing specific features of the trilobites that contribute to greater visual realism. The targeted adjustments in the prompts post-optimization are pivotal in directing the model to produce outputs that not only adhere more closely to the reference images but also showcase more pronounced and authentic trilobite characteristics. This approach demonstrates the model's capability to adapt and refine its output by learning from preferences, ultimately yielding higher match scores and visually richer renderings.

B. Qualitative Results II: Smoothness

Fig. 4 displays a series of frames. The sequences before and after preference optimization are shown. In the initial frames before optimization, the trilobite's movement appears somewhat jerky, especially its antennae. The corresponding prompt focuses on the trilobite's glide through the landscape. After preference optimization, the frames show a noticeable improvement in the fluidity of the trilobite's movement. The animation becomes smoother, with the trilobite seamlessly integrating into the motion of the surrounding plants. This creates a more naturalistic and visually appealing scene. The cause of this change is that the optimized prompt adds some frames and emphasizes the smooth, effortless glide of the trilobite and its streamlined body, highlighting how these characteristics should be reflected in the animation. This directive likely influenced the rendering process to focus on creating a smoother and more coherent movement pattern.

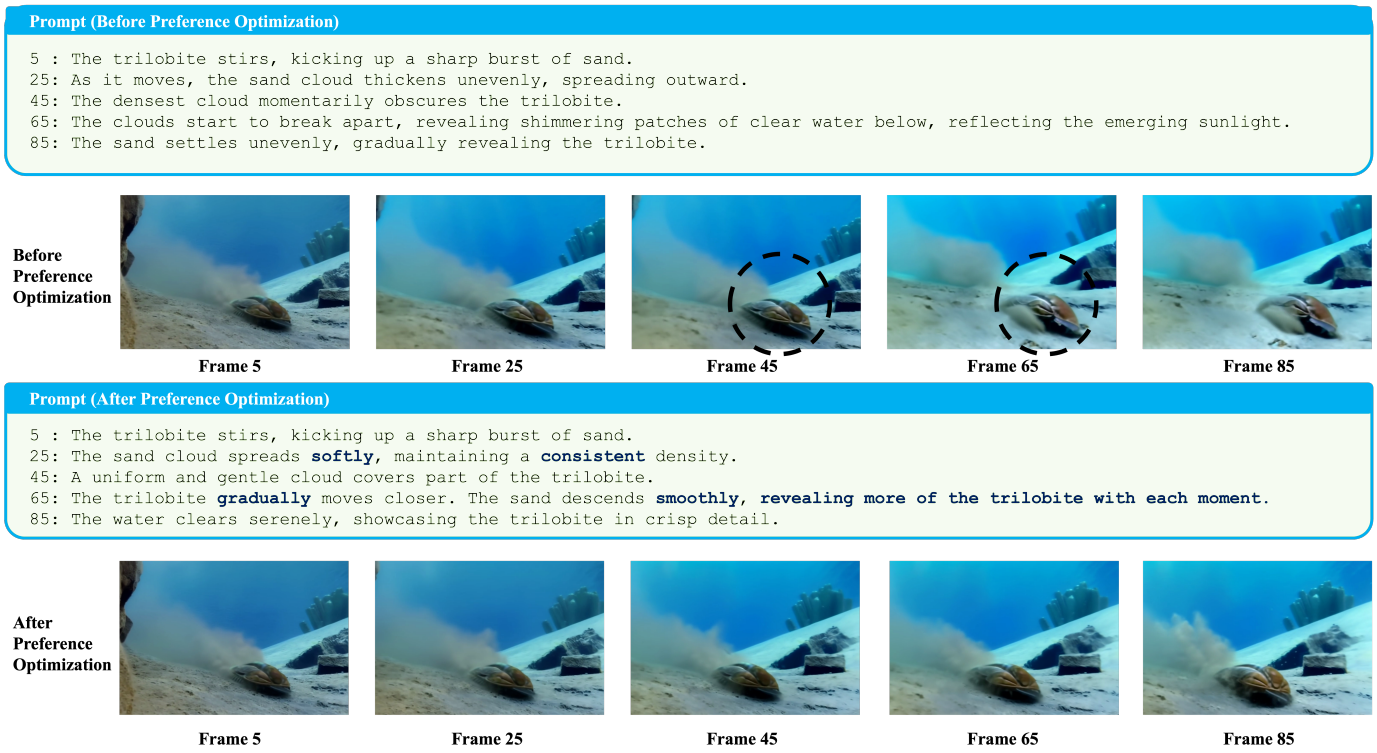

Fig. 5: Another qualitative comparison before and after SCRIPT WRITER preference optimization regarding video smoothness and continuity. The SCRIPT WRITER learns to **use words that indicate degree and process** to enhance the resulting video smoothness.

Fig. 5 gives another example where the SCRIPT WRITER learns to add some words to enhance the video smoothness. The comparison clearly demonstrates that the changes in the prompts, post-optimization, lead to significant improvements in the smoothness of the video.

C. Quantitative Results

We conduct quantitative comparisons to further evaluate our method.

Smoothness after KTO prompt training. Fig. 6 illustrates the Fréchet Inception Distance (FID) scores between adjacent frames in a generated video sequence, comparing results before and after preference optimization.

Before preference optimization, the blue line shows several peaks, particularly noticeable around frames 15, and 60-80, suggesting that the transitions between these frames are less smooth, with more noticeable visual discrepancies. After preference optimization, the dark line generally maintains lower FID scores throughout the sequence, with fewer and lower peaks compared to the blue line. This indicates that after optimization, the frames have greater visual consistency, and the transitions between them are smoother.

The overall trend in the graph demonstrates that preference optimization effectively reduces the FID scores across the majority of the video sequence. This improvement signifies that the video has become smoother post-optimization, with more consistent and visually coherent transitions between frames.

User study. We generate videos using four different methods and conduct a user study to evaluate their performance. Participants are asked to rate the videos on three criteria: smoothness, visual realism, and consistency with the prompt. We recruit participants for this study, each rating the videos on a scale from 1 to 4, where 4 indicates the highest possible score. This scoring system is equivalent to Average User Ranking (AUR), with higher scores indicating superior performance across the evaluated metrics.

Our method outperforms the other three methods in all evaluation criteria, indicating a significant improvement in video generation quality. This is evident from the higher scores across all three categories, confirming the effectiveness of our approach in producing videos that are smooth, visually realistic, and consistent with the given prompts. Particularly, the much higher scores regarding consistency with prompts achieved by our method highlight the effectiveness of our prompt learning method.

VI. CONCLUSION

In conclusion, our study leverages advanced generative AI techniques to address the challenges of reconstructing trilobite behavior from fossil records. By integrating computational methods with paleontological research, we demonstrate the potential to enhance our understanding of these ancient

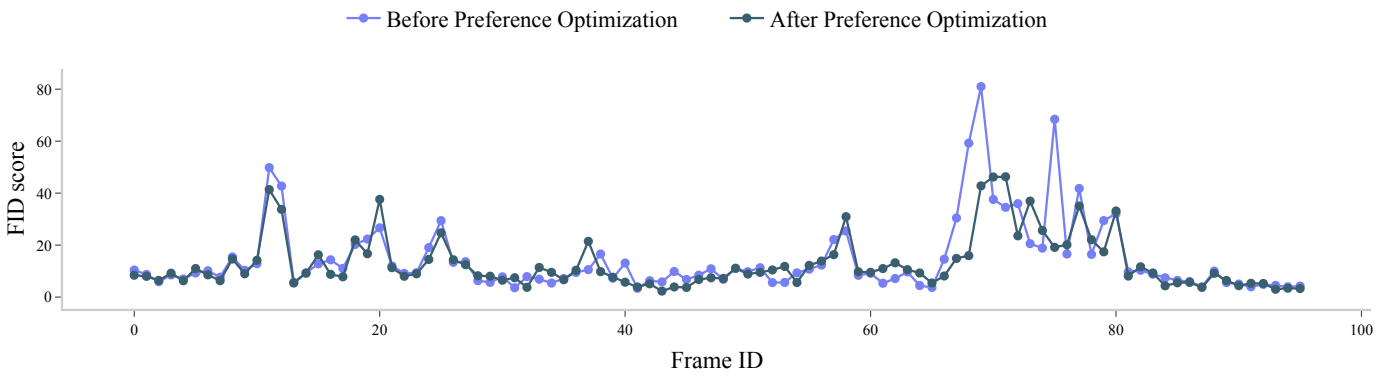

Fig. 6: Quantitative comparison: FID between adjacent frames. The x-axis represents the frame ID, ranging from 0 to 100, corresponding to the sequence of frames in the video. The y-axis quantifies the FID score, where a lower score indicates greater visual similarity and consistency between frames. The result shows that SCRIPT WRITER preference optimization effectively improves the smoothness of the generated video.

	Smoothness	Visual Realism	Consistency with prompt
Ours	**3.41± 0.21**	**3.70± 0.15**	**3.56± 0.27**
Gen3	3.02± 0.19	2.22± 0.25	2.41± 0.12
Pika	2.47 ± 0.25	2.67 ± 0.26	2.58 ± 0.14
Animatediff	1.27± 0.29	1.41± 0.23	1.44± 0.17

TABLE I: Quantitative comparison based on user study (mean ± var). A higher score indicates better performance.

creatures. Our proposed video generation framework, which incorporates realism and smoothness assessments into the workflow, produces more accurate and dynamic visualizations of trilobite movements. These enhanced animations not only improve scientific insights but also make the prehistoric world more accessible to the public. This interdisciplinary approach marks an advancement in both the fields of paleontology and (multi-modal) artificial intelligence, opening new avenues for future research and educational opportunities.

ACKNOWLEDGMENT

We deeply appreciate Qiang Ou (China University of Geosciences, Beijing), Degan Shu (Northwest University, Xi' an), Jian Han (Northwest University, Xi' an), Meirong Cheng (Northwest University, Xi' an) for generously providing the image of trilobites that supported this study.

REFERENCES

[1] A. El Albani, A. Mazurier, G. D. Edgecombe, A. Azizi, A. El Bakhouch, H. O. Berks, E. H. Bouougri, I. Chraiki, P. C. Donoghue, C. Fontaine *et al.*, "Rapid volcanic ash entombment reveals the 3d anatomy of cambrian trilobites," *Science*, vol. 384, no. 6703, pp. 1429–1435, 2024.

[2] R. Fortey, "The palaeoecology of trilobites," *Journal of zoology*, vol. 292, no. 4, pp. 250–259, 2014.

[3] C. trilobite fossil from Utah, "Trilobites and end of cambrian explosion," *PNAS*, vol. 116, no. 10, pp. 3935–3937, 2019.

[4] J. Bergström, "Organization, life, and systematics of trilobites," in *Organization, life, and systematics of trilobites*, 1973, pp. 1–69.

[5] N. C. Hughes, "The evolution of trilobite body patterning," *Annu. Rev. Earth Planet. Sci.*, vol. 35, pp. 401–434, 2007.

[6] R. Levi-Setti, *Trilobites*. University of Chicago Press, 1995.

[7] Q. Ou, D. Shu, J. Han, X. Zhang, Z. Zhang, and J. Liu, "A juvenile redlichiid trilobite caught on the move: Evidence from the cambrian (series 2) chengjiang lagerstatte, southwestern china," *Palaios*, vol. 24, no. 7, pp. 473–477, 2009.

[8] M. J. Hopkins, "Development, trait evolution, and the evolution of development in trilobites," *Integrative and Comparative Biology*, vol. 57, no. 3, pp. 488–498, 2017.

[9] E. B. Hunt, *Artificial intelligence*. Academic Press, 2014.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," in *The Eleventh International Conference on Learning Representations*, 2022.

[12] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 954–15 964.

[13] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7623–7633.

[14] A. Sohrabi, A. Kadkhodaie, and R. Kadkhodaie-Ilkhchi, "Artificial intelligence approach to palaeogeography and evolutionary trend analysis of laurentian brachiopod fauna in the rhynchotrema-hiscobeccus lineage," *Palaeogeography, Palaeoclimatology, Palaeoecology*, vol. 562, p. 110114, 2021.

[15] Y. Li, M. Min, D. Shen, D. Carlson, and L. Carin, "Video generation from text," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[16] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[17] W. Ren, H. Yang, G. Zhang, C. Wei, X. Du, S. Huang, and W. Chen, "Consisti2v: Enhancing visual consistency for image-to-video generation," *arXiv preprint arXiv:2402.04324*, 2024.

[18] R. Henschel, L. Khachatryan, D. Hayrapetyan, H. Poghosyan, V. Tadevosyan, Z. Wang, S. Navasardyan, and H. Shi, "Streamingt2v: Consistent, dynamic, and extendable long video generation from text," *arXiv preprint arXiv:2403.14773*, 2024.

[19] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[20] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[21] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet,

"Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.
[22] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.
[23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[24] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.
[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
[26] N. Bouali and V. Cavalli-Sforza, "A review of text-to-animation systems," *IEEE Access*, 2023.
[27] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.
[28] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163.
[29] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[30] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela, "Kto: Model alignment as prospect theoretic optimization," 2024. [Online]. Available: https://arxiv.org/abs/2402.01306
[31] P. Lab, "Pika lab," https://pika.art/, 2024.
[32] Gen-2, "Gen-2: The next step forward for generative ai," https://research.runwayml.com/gen2, 2024.
[33] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.
[34] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.
[35] Y. Wang, Q. Liu, and C. Jin, "Is rlhf more difficult than standard rl? a theoretical perspective," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[36] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 835–10 866.
[37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[38] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and uncertainty*, vol. 5, pp. 297–323, 1992.
[39] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.
[40] C. Wang, Y. Deng, Z. Lv, S. Yan, and A. Bo, "Q*: Improving multi-step reasoning for llms with deliberative planning," *arXiv preprint arXiv:2406.14283*, 2024.
[41] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu *et al.*, "Deepseek llm: Scaling open-source language models with longtermism," *arXiv preprint arXiv:2401.02954*, 2024.
[42] T. Wei, L. Zhao, L. Zhang, B. Zhu, L. Wang, H. Yang, B. Li, C. Cheng, W. Lü, R. Hu *et al.*, "Skywork: A more open bilingual foundation model," *arXiv preprint arXiv:2310.19341*, 2023.
[43] T. Wei, B. Zhu, L. Zhao, C. Cheng, B. Li, W. Lü, P. Cheng, J. Zhang, X. Zhang, L. Zeng *et al.*, "Skywork-moe: A deep dive into training techniques for mixture-of-experts language models," *arXiv preprint arXiv:2406.06563*, 2024.
[44] K. Lu, H. Yuan, Z. Yuan, R. Lin, J. Lin, C. Tan, C. Zhou, and J. Zhou, "# instag: Instruction tagging for analyzing supervised fine-tuning of large language models," in *The Twelfth International Conference on Learning Representations*, 2023.
[45] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.
[46] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," *Advances in neural information processing systems*, vol. 32, 2019.
[47] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 1415–1428, 2021.
[48] X. Shen, X. Li, and M. Elhoseiny, "Mostgan-v: Video generation with temporal motion styles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5652–5661.
[49] W. Myers', "Phacops trilobites art print," https://fineartamerica.com/featured/phacops-trilobites-walter-myers.html?product=art-print, phacops Trilobites Art Print.
[50] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, "The role of imagenet classes in fr\'echet inception distance," *arXiv preprint arXiv:2203.06026*, 2022.
[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
[52] P. Aglave and V. S. Kolkure, "Implementation of high performance feature extraction method using oriented fast and rotated brief algorithm," *Int. J. Res. Eng. Technol*, vol. 4, pp. 394–397, 2015.
[53] N. Antony and B. R. Devassy, "Implementation of image/video copy-move forgery detection using brute-force matching," in *2018 2nd International conference on trends in electronics and informatics (ICOEI)*. IEEE, 2018, pp. 1085–1090.
[54] F. K. Noble, "Comparison of opencv's feature detectors and feature matchers," in *2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*. IEEE, 2016, pp. 1–6.
[55] A. S. Photo, "2hg309h," https://www.alamy.com/an-illustration-of-a-trilobite-moving-about-on-a-cambrian-period-400-million-years-ago-sea-bottom-trilobites-are-a-well-known-fossil-group-image457370189.html, 2016, an illustration of a Trilobite moving about on a Cambrian Period (400 million years ago) sea bottom. Trilobites are a well-known fossil group of extinct marine arthropods that form the class Trilobita.